

New Trends in IT

Referat zum Thema
„Suchmaschine Lucene“

Gruppe
Stefan Habel, Thomas Vogel

Inhalt

- Was ist Lucene?
- Geschichte
- Grundsätzliche Vorgehensweise
- Unterstützte Dateiformate
- Suchtypen
- Portierungen in andere Programmiersprachen
- Projekte/Software die Lucene verwenden

Was ist Lucene?

- Suchmaschinen-Framework zur Volltextsuche
- Erstellt in JAVA
- Keine fertige Suchmaschine eher eine grobe Architektur
- Stellt Klassen und Funktionen zur Verfügung
- Detailspekte durch Entwickler zu implementieren

Geschichte

- Erfinder Doug Cutting
- Verantwortlich für die Suchhilfe „Sherlock“ von Apple
- Erste Idee 1997, erster Prototyp 1998
- Der Name ist der Vorname von Doug Cutting's Ehefrau
- 2001 Umzug des Lucene Projekt von Sourceforge zur Apache Foundation
- Seit 2005 Top-Level-Apache Projekt

Grundsätzliche Vorgehensweise

- Der zu durchsuchende Text muss komplett vorliegen
- Durchsuchen des gesamten Textes
- Erstellen von Indizes
- Index → Dokumente → Felder → Ausdrücke → Worte
- Bei Suchanfrage durchsuchen der Indizes

Unterstützte Dateiformate

- Grundsätzlich alle Arten von Dateien und Formaten
- Implementiert ist nur die Suche in Standard-Text Dateien im ASCII Format
- Für HTML, DOC, PDF etc. müssen vom Entwickler Filter zur Verfügung gestellt werden
- Projekte die solche Filter zur Verfügung stellen sind zum Beispiel:
 - PDFBox, Jtidy oder Apache POI

Suchtypen

- Lucene unterstützt unter anderem folgende Suchtypen:
 - Termsuche (Suche nach Sätzen und Phrasen)
 - Boolesche Suche
 - Wildcardsuche (*.*, xyz* , „?“)
 - Fuzzy- oder Proximity-Suche (Undeutliche Suche, Annäherungssuche)

Portierungen

Des weiteren existieren Lucene Adaptionen in anderen Programmiersprachen:

- Perl → Plucene
- Python → Pylucene
- Ruby → Ferret
- C++ → CLucene
- C# → DotLucene

Projekte die Lucene verwenden

Bei Lucene handelt es sich nicht nur um eine theoretische Idee, produktiv im Einsatz zum Beispiel bei:

- Desktop Suche „Beagle“ C#
- Desktop Suche „Strigi“ C++
- Vollständige Suchmaschine „Nutch“ im Einsatz bei Wikipedia
- WebGate Anywhere ein CMS das Lucene verwendet